State of Oregon
Department of Environmental Quality Memorandum

To: **DEQ Cleanup Program PMT** **Date:** 5 September 2006
Sally Puent, Manager, NWR-WQ
Gene Foster, Manager, DEQ- Lab

From: **Mike Poulsen**
Cleanup and Lower Willamette Section
Keith Johnson
Manager, CU and Lower Willamette Section

Subject: Evaluation of Reliability of Potential Freshwater Sediment Screening Values

The Regional Sediment Evaluation Team (RSET) and DEQ desire to have uniform freshwater sediment screening values that will determine what actions may be necessary at both sediment dredge sites and sediment cleanup sites. Assuming broader consensus in the RSET group, the screening values will also be suitable for consistent use by the federal agencies (EPA, NMFS, and NOAA), and also by the sediment dredging and cleanup programs of Oregon and the two other northwestern states (Washington and Idaho).

The purpose of this memo is to identify key aspects of the proposed methodology, the technical concepts involved, and the implications to DEQ should we move forward in adopting this approach.

Tiered Screening Levels Based on Toxicity

RSET has proposed to use draft un-promulgated sediment quality guidelines (SQGs) developed by Dr. Teresa Michelsen for Washington state in 2002. Teresa used a floating percentile method to derive the draft SQGs.

Two screening levels are being proposed, based on different definitions of toxicity ("hit").

- SL1 values are based on a 10% difference (*e.g.*, in mortality or another toxic endpoint such as growth) in a bioassay result compared with the control (a "clean" sample).
- SL2 values are based on a 25% difference from control, indicating a higher degree of toxicity.

In the traditional three tier system of screening, *SL1 values will be used to screen out sediment samples as non-toxic*. *SL2 values will be used to screen in samples as toxic*. Samples with concentrations *between SL1 and SL2 values will require additional evaluation and other lines of evidence*.

Evaluating the Reliability of Screening Levels- Discussion

Another consideration in the determination of the screening values is how accurate the methodology predicts proper results. Two key reliability measures are the following:

- False negative – the percentage of known toxic samples that are incorrectly screened out using specified screening values

- False positive – the percentage of known non-toxic samples that are incorrectly screened in

Table 1 shows the definitions of all the reliability measures, as well as the results of Teresa's calculations. The SQG subcommittee is recommending that we use the screening values developed from a false negative rate of 15%. Figure 1 shows a representation of the results for SL1 using the false negative rate of 15%. Given our definition of a hit, all samples must be placed in one of four bins, either a correctly predicted hit, a correctly predicted no-hit, a hit incorrectly predicted as a no-hit, or a no-hit incorrectly predicted as a hit. In Figure 1, the number of correctly predicted hits is 34. Dividing by the total number of hits (40) gives a sensitivity of 85%. The corresponding false negative rate is 15% (= 100% - 85%). In other words, using the SL1 screening values, we will miss only 15% of the samples known to be toxic. This is a reasonably good result, and this measure should be a primary focus of the agencies.

However, the DEQ and the other agencies should also be interested in how reliably we predict no-hits (measured by predicted-no-hit efficiency). Based on our definition of a hit and the proposed screening criteria, Figure 1 shows that we are 67% confident that a sample predicted to be a no-hit at the SL1 screening level will in fact be non-toxic if we were to conduct a bioassay. *Stated another way, in the existing dataset, one third of the samples predicted as no-hit were toxic in a bioassay.*

The above results give different perspectives on the reliability of the screening values. On one hand, we are reasonably confident that we will screen in known toxic samples at the SL1 level. *On the other hand, for samples screened out as non-toxic, there is still a good chance that they may still be toxic*.

RSET has not established criteria for making a decision regarding the acceptability of these reliability results. We expect that Oregon DEQ managers will likely accept a false negative rate of 15%. However, Jennifer Peterson and Mike Poulsen have concerns about a false-predicted-no-hit rate of 33%, particularly if the screening values are used as a sole line of evidence. In addition to the specific numeric measures of reliability calculated for the model, we are concerned that the current model is based on a limited dataset primarily from western Washington and western Oregon, and has not undergone validation.

For these reasons, the SQG subcommittee agreed to state in the guidance that regulatory agencies may require additional evaluations (possibly including bioassays) even if concentrations of chemicals in sediment are below SL1 screening values. It is likely that this condition will be proposed only for the cleanup programs. This appears acceptable to us for the Cleanup program, and is similar to our current approach for evaluating sediments. We do not know if this is acceptable to the dredging program.

Reliability results differ with the selected level of toxicity (SL1 or SL2). At the level of toxicity used to develop the SL2 screening values, the predicted-no-hit rate increases from 67% (for SL1) to about 84% (see Table 1), with a correspondingly lower false-predicted-no-hit rate (16%) that DEQ managers may consider acceptable. For a sample with concentrations below SL2

screening levels, there is only a 1/6 chance of the sediment being toxic at the SL2 higher level of toxicity.

In addition, the regulated community will rightly be concerned about the other reliability measures (e.g., false positives and false predicted hits). One of the great benefits of the floating percentile method is the ability to optimize screening values by reducing false positives for a given false negative rate. However, RSET has not established explicit criteria for the acceptability of these rates. Jennifer and Mike consider it more relevant to attempt to optimize the false positive and related rates in developing upper-tier screening values (e.g., SL2) that are more indicative of sediments requiring remediation, rather than optimizing these rates for the lower-tier screening values (SL1).

In discussions with the RSET SQG subcommittee, Jennifer and Mike proposed that more conservative screening values, such as TELs, be used for the lower screening values. TEL values have more optimal false negative and false-predicted-no-hit rates for a lower screen. The reliability estimates for various screening approaches are shown in Table 2. Actual chemical concentration values for the different screening approaches are shown in Table 3.

It was made clear to Jennifer and Mike that RSET would not accept TEL or similar values as lower-tier screening values for dredging decisions. (We do not know NOAA's opinion. They are part of the SQG committee, but not the subcommittee of Oregon and Washington that the USACE has recently convened.) The USACE considers the false positive rates too high, and is concerned that very few sediment areas will be screened out if values such as TELs are used for the lower screen. The proposed compromise is to allow states and other regulatory agencies the option of requiring additional evaluation (including bioassays) for samples with concentrations below SL1 screening values.

Jennifer and Mike think that the proposal will be workable for the Cleanup program, but we are unclear about the implications for the dredging program. It may be appropriate for the Cleanup program and the dredge program to have different evaluation approaches to applying the SL1 screening values. However, we consider it appropriate to consider additional evaluation of large dredge sites, and sites in Eastern Region (which was not adequately represented in the database used to derive the screening values). Additional data (primarily bioassay test results) from both cleanup and dredging sites can be used in the future to refine the development of freshwater screening values. The screening values for marine sites have received more scrutiny, and are considered more reliable. Jennifer and Mike did not review the marine screening values.

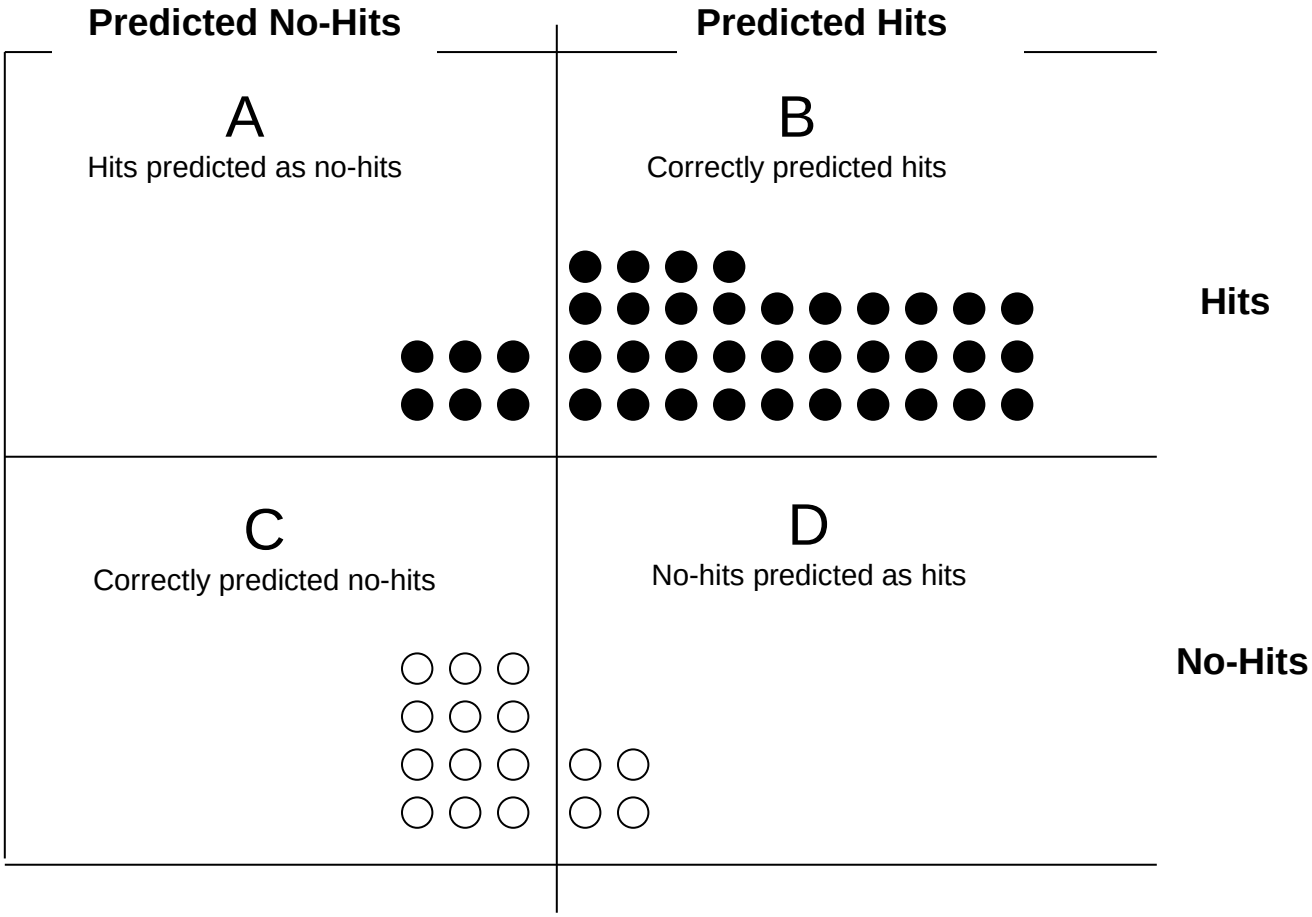
Summary

Here is the how the draft SEF is expected to address toxicity screening.

- Freshwater sediment concentrations <u>below the SL1 screening values</u>
    - Cleanup Program – Responses are flexible. We propose that ODEQ use the SL1 values as a line of evidence, but use other values, such as TELs, as a more

reliable lower screen. Washington will also use other lines of evidence until the screening values are validated.
    - Dredging Program – RSET will likely propose that the dredge programs use the SL1 values as a definitive screen. It is unclear if SL1 values will be acceptable as a definitive screen in Oregon. Most inwater disposal of sediments will be in marine waters, so this may not be a frequent issue.
- Freshwater sediment concentrations between the SL1 and SL2 screening values
    - Cleanup Program – Additional lines of evidence (such as bioassays) will be used to determine if the sediment toxicity warrants further action.
    - Dredging Program -- Additional lines of evidence (such as bioassays) will be used to determine if the sediment is unacceptable.
- Freshwater sediment concentrations above the SL2 screening values
    - Cleanup Program – Sediment is presumed to be toxic. Bioassays could be used to show that the sediment is non-toxic.
    - Dredging Program -- Sediment is presumed to be toxic. Bioassays could be used to show that the sediment is non-toxic.

Figure 1.
Reliability Measures of Proposed SL1 Screening Criteria



Sensitivity = B / (A + B) = 0.85
False Negatives = A / (A + B) = 0.15

Efficiency = C / (C + D) = 0.75
False Positives = D / (C + D) = 0.25

Predicted-Hit Efficiency = B / (B + D) = 0.89
False Predicted Hits = D / (B + D) = 0.11

Predicted-No-Hit Efficiency = C / (A + C) = 0.67
False Predicted No-Hits = A / (A + C) = 0.33

This page left intentionally blank.

Table 1
Reliability Estimates for Proposed Freshwater Sediment Screening Values in Draft SEF

		Percentage (%)	
	Screening Level:	SL1	SL2
Reliability Measure	**Definition**		
Sensitivity (Hit Efficiency)	Percentage of known toxic samples that are correctly screened in	84	85
False Negative	Percentage of known toxic samples that are incorrectly screened out	16	15
(No-Hit) Efficiency	Percentage of known non-toxic samples that are correctly screened out	75	75
False Positive	Percentage of known non-toxic samples that are incorrectly screened in	25	25
Predicted-Hit Efficiency	Percentage of screened-in samples that are toxic	88	77
False Predicted Hit	Percentage of screened-in samples that are non-toxic	12	23
Predicted-No-Hit Efficiency	Percentage of screened-out samples that are non-toxic	67	84
False Predicted-No-Hit	Percentage of screened-out samples that are toxic	33	16

Note:
See Figure 1 for a graphical presentation of the reliability measures for SL1.

Table 2
Comparison of Reliability Estimates for Various Screening Values

		Lower Screen[a]				Upper Screen[a]			
Reliability Measure	**Definition**	**SL1**	**TEL**	**TEC**	**LEL**	**SL2**	**PEL**	**PEC**	**SEL**
Sensitivity (Hit Efficiency)	Percentage of known toxic samples that are correctly screened in	84	96	87	95	85	70	62	58
False Negative	Percentage of known toxic samples that are incorrectly screened out	16	4	13	5	15	30	38	42
(No-Hit) Efficiency	Percentage of known non-toxic samples that are correctly screened out	75	13	22	18	75	49	60	69
False Positive	Percentage of known non-toxic samples that are incorrectly screened in	25	87	78	82	25	51	40	31
Predicted-Hit Efficiency	Percentage of screened-in samples that are toxic	88	49	49	50	77	36	39	44
False Predicted Hit	Percentage of screened-in samples that are non-toxic	12	51	51	50	23	64	61	56
Predicted-No-Hit Efficiency[b]	Percentage of screened-out samples that are non-toxic	67	79	66	81	84	80	79	80
False Predicted-No-Hit	Percentage of screened-out samples that are toxic	33	21	34	19	16	20	21	20

Notes:

a) From Development of Freshwater Sediment Quality Values for Use in Washington State, Phase I Task 6 Report, Sept. 2002, Table 3-3.
TEL = Threshold Effects Level PEL = Probable Effects Level
TEC = Threshold Effects Concentration PEC =Probably Effects Concentration
LEL = Lowest Effect Level SEL = Severe Effect Level

b) Predicted-No-Hit Efficiency, PNHE = (Eff/FP)(FPH/PHE)(Sen/FN) / [(Eff/FP)(FPH/PHE)(Sen/FN) + 1]

Table 3
Comparison of Proposed RSET Freshwater Sediment Screening Values With Other Screening Values

	Proposed RSET Screening Levels[a]		Other Freshwater Sediment Values[b,c]				
Chemical	**SL1**	**SL2**	**TEL**	**TEC**	**PEL**	**PEC**	**AET**
Metals (mg/kg)							
Antimony	0.4	0.6					64
Arsenic	20	51	5.9	9.8	17	33	40
Cadmium	0.6	1	0.6	0.99	3.5	4.5	7.6
Chromium	95	100	37	43	90	110	280
Copper	80	830	36	32	200	150	840
Lead	335	430	35	36	91	130	260
Mercury	0.5	0.75	0.17	0.18	0.49	1.1	0.56
Nickel	60	70	18	23	36	49	46
Silver	2	2.5					4.5
Zinc	140	160	120	120	320	460	520
Tributyltin	75	75					
SVOCs (ug/kg)							
Total PCBs	60	120	34	60	280	680	21
DEHP	230	320					750
Butylbenzylphthalate	260	370					
Di-n-butylphthalate							
Dibenzofuran	400	440					32,000
Pesticides (ug/kg)							
Total DDTs			1.2	5.3	4.8	570	
PAHs (ug/kg)							
Total LPAH	6,600	9,200					74,000
Total HPAH	31,000	54,800					91,000
Total PAHs				1,600		23,000	170,000
Acenaphthene	1,060	1,320	6.7		89		4,100
Acenaphthylene	470	640	5.9		130		2,200
Anthracene	1,200	1,580	47	57	250	850	2,800
Benz[a]anthracene	4,260	5,800	32	110	390	1,100	7,700
Benzo[a]pyrene	3,300	4,810	32	150	780	1,500	11,000
Benzo[g,h,i]perylene	4,020	5,200					1,400
Chrysene	5,940	6,400					
Dibenz[a,h]anthracene	800	840	6.2	33	140		230
Fluoranthene	11,000	15,000	110	420	2,400	2,200	21,000
Fluorene	1,000	3,000	21	77	140	540	4,200
Naphthalene	500	1,310	35	180	390	560	46,000
Phenanthrene	6,100	7,600	42	200	520	1,200	15,000
Pyrene	8,800	16,000	53	200	880	200	23,000

Notes:

a) Draft Sediment Evaluation Framework, Sept. 2005, Table 7-1.

b) Development of Freshwater Sediment Quality Values for Use in Washington State, Phase I Task 6 Report, Sept. 2002, Appendix H.

c) SL1 = Screening Level 1
SL2 = Screening Level 2
TEL = Threshold Effects Level
TEC = Threshold Effects Concentration
PEL = Probable Effects Level
PEC = Probably Effects Concentration
AET = Apparent Effects Threshold